

Semiparametric Adaptive Estimation With Nonignorable Nonresponse Data

Kosuke Morikawa*

Graduate school of Engineering Science, Osaka University
and

Jae Kwang Kim

Department of Statistics, Iowa State University

December 30, 2016

Abstract

When the response mechanism is believed to be nonignorable or not missing at random (NMAR), a valid analysis requires stronger assumptions about the data than do standard statistical methods. Semiparametric estimators have been developed under the correct model specification assumption for the response mechanism. In this paper, we consider a scheme for obtaining the optimal estimation for the parameters such as the mean and propose two semiparametric adaptive estimators that do not require any model assumptions except for the response mechanism. Asymptotic properties of proposed estimators are discussed, and we present an application to Korean Labor and Income Panel Survey data.

Keywords: Not missing at random (NMAR); Incomplete data; Estimating functions; Semiparametric Estimation.

*The research of the first author is supported by a grant from JSPS KAKENHI (Grant Number JP16J03875) and the research of the second author is partially supported by a grant from NSF (MMS-1324922).

1 Introduction

Handling missing data often requires some assumptions about the response mechanism. If the study variable does not affect the probability of the response, the response mechanism is called ignorable. If, on the other hand, the response probability of a study variable depends on that variable directly, the response mechanism is called nonignorable (Little & Rubin 2002). A nonignorable nonresponse is more troublesome than an ignorable one, because the response probability cannot be verified using the observed study variables only. Thus, to estimate the parameters in the response model under a nonignorable nonresponse, additional assumptions about the study variable are often required.

Let r be the response indicator of the study variable y with auxiliary variable x , where r takes 1 if y is observed, and takes 0 otherwise. An ignorable nonresponse can be understood as the conditional independence of r and y given \mathbf{x} , namely $r \perp y \mid \mathbf{x}$. Diggle & Kenward (1994) proposed a fully parametric approach to analyze nonignorable nonresponse data; their method requires two parametric models: (i) an outcome model, $y \mid \mathbf{x}$; and (ii) a response model, $r \mid \mathbf{x}, y$. In practice, it is difficult to verify models (i) and (ii), because some of Y are not observed. In the fully parametric approach, model identification and model misspecification can be a problem, and sensitivity analysis has been recommended (Scharfstein et al. 1999, Rotnitzky et al. 2001, Verbeke et al. 2001, Tsiatis 2006). Sverchkov (2008) and Riddles et al. (2016) proposed a fully parametric approach that uses different model specifications based on (i) $y \mid \mathbf{x}, r = 1$, and (ii) $r \mid \mathbf{x}, y$. Their approach is more practical, because one can verify a model for $[y \mid \mathbf{x}, r = 1]$ from the responses; however, because it is a fully parametric approach, it is still subject to model misspecification.

Recently, some semiparametric approaches have been proposed for nonignorable non-responses. Tang et al. (2003) proposed a maximum pseudolikelihood estimator that does

not require specification of the response mechanism, and a distinct condition is used to identify each model. Zhao & Shao (2015) extended the method of Tang et al. (2003) to generalized linear models, and their method does not require assumptions about the response model. Fitzmaurice et al. (2005) and Skrondal & Rabe-Hesketh (2014) proposed protective estimators that do not require specification of the response mechanism, but the application of this approach is limited to situations in which Y is binary. Kim & Yu (2011) proposed a semiparametric method for estimating $E(Y)$ using a parametric response model, but a validation sample is required in order to estimate the parameters in the response model. Tang et al. (2014) used the method of empirical likelihood to extend the method of Kim & Yu (2011) to estimate more-general parameters. In Tang et al. (2016), the method of Qin et al. (2002) was used to construct a $n^{1/2}$ -consistent estimator without a validation sample. Morikawa et al. (2016) used the kernel regression estimator to remove the parametric model assumption on model (i), $y \mid \boldsymbol{x}, r = 1$. Chang & Kott (2008) and Wang et al. (2014) considered a generalized method of moments (GMM) estimator that uses only a response model, but their method is generally less efficient than the likelihood-based one. Recently, Shao & Wang (2016) proposed a semiparametric inverse propensity weighting method using the nonresponse instrumental variable assumption of Wang et al. (2014). Miao & Tchetgen Tchetgen (2016) considered doubly robust estimators under a nonignorable nonresponse. Their estimator has consistency and asymptotic normality, if the conditional odds ratio model is true and if either the baseline outcome density model or the baseline response model is true. Thus, we note that in general, the doubly robust estimators require that two out of the three models be correctly specified.

In this paper we assume a parametric model for $[r \mid \boldsymbol{x}, y]$ and give an optimal estimator for the model parameters describing the response mechanism. We derive the semiparametric efficiency bound (Tsiatis 2006) and propose two semiparametric estimators that each

attains the lower bound. The first estimator does not require any additional assumptions, but it is more efficient than the approach of Morikawa et al. (2016). The second is an adaptive estimator, such as the generalized linear estimator proposed by Liang & Zeger (1986); although it requires specification of the model of $[y \mid \boldsymbol{x}, r = 1]$, it has consistency and asymptotic normality, even if that specification is wrong. Furthermore, if the model specification is correct, it attains the lower bound.

2 Basic Setup

Let (\boldsymbol{z}_i, r_i) , $i = 1, \dots, n$ be n realizations from a joint distribution of (\boldsymbol{Z}, R) , where $\boldsymbol{Z} = (\boldsymbol{X}^\top, Y)^\top$, \boldsymbol{X} is a covariate vector, Y is a response variable, and R is a response indicator of Y , i.e., it takes 1 if Y is observed, and takes 0 otherwise. Also, let $\boldsymbol{G}_R(\boldsymbol{Z})$ be the observed data when the response indicator is R , i.e., $\boldsymbol{G}_1(\boldsymbol{Z}) = \boldsymbol{Z}$ and $\boldsymbol{G}_0(\boldsymbol{Z}) = \boldsymbol{X}$. Suppose that a response model is $\pi(\boldsymbol{z}; \boldsymbol{\phi})$ with a q -dimensional parameter $\boldsymbol{\phi} \in \Phi$. For model identification, we assume that there exists \boldsymbol{X}_1 in $\boldsymbol{X} = (\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top)^\top$ such that \boldsymbol{X}_2 is independent of R , given \boldsymbol{X}_1 and Y . Such variable is referred to as nonresponse instrumental variable (Wang et al. 2014). For the parametric model for π , we can consider several possibilities, such as a logistic model, a probit model, or a complementary log-log nonignorable model (Kim & Yu 2011). Also, let $\theta \in \Theta$ be a one-dimensional parameter such that $\theta = E\{U(Z)\}$, where U is a known function. For example, if we are interested in $E(Y)$ and $P(Y > a)$, then U is Y and $I(Y > a)$, respectively, where I is an indicator function and a is an arbitrary real number.

In this paper, we consider semiparametric estimation of $(\boldsymbol{\phi}, \theta)$. In particular, we find the most efficient estimator among the regular asymptotically linear (RAL) estimators (Bickel et al. 1998, Robins et al. 1994, Tsiatis 2006). An estimator in a semiparametric

model is called regular if it is regular in every regular parametric submodel. A finite-dimensional estimator $\hat{\boldsymbol{\xi}}$ in a parametric submodel is called regular if the data are generated by $f(\mathbf{Z}, \mathbf{G}_R(\mathbf{Z}); \boldsymbol{\xi}_n)$, where, if $\boldsymbol{\xi}_n$ is a finite-dimensional parameter and $n^{1/2}(\boldsymbol{\xi}_n - \boldsymbol{\xi}^*)$ converges to a constant, then $n^{1/2}(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}_n)$ converges uniformly to its limiting distribution. Also, $(\hat{\boldsymbol{\phi}}^\top, \theta)^\top$ is said to be an asymptotically linear estimator of $(\boldsymbol{\phi}_0^\top, \theta_0)^\top$ if

$$n^{1/2}((\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)^\top, \hat{\theta} - \theta_0)^\top = n^{-1/2} \sum_{i=1}^n \boldsymbol{\varphi}(R_i, \mathbf{G}_{R_i}(\mathbf{Z}_i); \boldsymbol{\phi}_0, \theta_0) + o_p(1), \quad (1)$$

where the function $\boldsymbol{\varphi}$ is called the influence function of $(\hat{\boldsymbol{\phi}}^\top, \hat{\theta})^\top$. In some cases, there exists a regular estimator, but not asymptotically linear. However, Hájek's representation theorem shows that most efficient regular estimators are asymptotically linear, and thus it is reasonable to discuss efficiency among RAL estimators (Tsiatis 2006).

Classical approaches for nonignorable nonresponse data are based on correct specification for the distribution $Y \mid X$ (Diggle & Kenward 1994). This requirement can be problematic because the specification cannot be verified under a nonignorable nonresponse. Chang & Kott (2008) proposed a semiparametric estimator for $\boldsymbol{\phi}$ that is based on the following estimating equations:

$$\sum_{i=1}^n \left(1 - \frac{r_i}{\pi(\mathbf{z}_i; \boldsymbol{\phi})} \right) \mathbf{g}(\mathbf{x}_i) = 0, \quad (2)$$

where $\mathbf{g} : \mathbb{R}^d \rightarrow \mathbb{R}^q$ is an arbitrary function of \mathbf{X} . Note that although this estimator has consistency and asymptotic normality under certain regularity conditions, its optimality is not guaranteed. If $q = d + 1$, a typical choice for \mathbf{g} is $\mathbf{g}(\mathbf{x}) = (1, \mathbf{x}^\top)^\top$. However, if $p > d + 1$, it is not clear how to determine the best \mathbf{g} in terms of minimizing the asymptotic variance, although Chang & Kott (2008) proposed some functions for \mathbf{g} .

Recently, Riddles et al. (2016) proposed a more efficient estimator that uses a parametric model for $Y \mid \mathbf{X}, R = 1$. Using the mean score theorem (Louis 1982), the maximum-

likelihood estimator can be obtained by solving

$$n^{-1} \sum_{i=1}^n [r_i \mathbf{s}_1(\mathbf{z}_i; \boldsymbol{\phi}) + (1 - r_i) E_0\{\mathbf{s}_0(\mathbf{Z}; \boldsymbol{\phi}) \mid \mathbf{x}_i\}] = 0, \quad (3)$$

where $\mathbf{s}_r(\mathbf{z}; \boldsymbol{\phi})$ is the score function of $\boldsymbol{\phi}$, that is,

$$\mathbf{s}_r(\mathbf{z}; \boldsymbol{\phi}) = \frac{\{r - \pi(\mathbf{z}; \boldsymbol{\phi})\} \dot{\pi}(\mathbf{z}; \boldsymbol{\phi})}{\pi(\mathbf{z}; \boldsymbol{\phi}) \{1 - \pi(\mathbf{z}; \boldsymbol{\phi})\}},$$

$\dot{\pi}(\mathbf{z}; \boldsymbol{\phi}) = \partial \pi(\mathbf{z}; \boldsymbol{\phi}) / \partial \boldsymbol{\phi}$, and $E_0(\cdot \mid \mathbf{x})$ is the conditional expectation of Y given \mathbf{X} and $R = 0$. To compute $E_0(\cdot \mid \mathbf{x})$, using Bayes' formula, Riddles et al. (2016) proposed using

$$n^{-1} \sum_{i=1}^n \left\{ r_i \mathbf{s}_1(\mathbf{z}_i; \boldsymbol{\phi}) + (1 - r_i) \frac{E_1(O(\mathbf{Z}; \boldsymbol{\phi}) \mathbf{s}_0(\mathbf{Z}; \boldsymbol{\phi}) \mid \mathbf{x}_i)}{E_1(O(\mathbf{Z}; \boldsymbol{\phi}) \mid \mathbf{x}_i)} \right\} = 0, \quad (4)$$

where $O(\mathbf{z}; \boldsymbol{\phi}) = \{1 - \pi(\mathbf{z}; \boldsymbol{\phi})\} / \pi(\mathbf{z}; \boldsymbol{\phi})$, and $E_1(\cdot \mid \mathbf{x})$ is the conditional expectation on Y given \mathbf{X} and $R = 1$. The conditional expectation is computed by assuming a parametric model $f_1(y \mid \mathbf{x}; \boldsymbol{\gamma}) = f(y \mid \mathbf{x}, r = 1; \boldsymbol{\gamma})$. This may increase the efficiency, however, misspecification of the f_1 model causes θ to be inconsistent.

3 Efficiency Bound

In this section, we provide an estimator for $(\boldsymbol{\phi}^\top, \theta)^\top$ that is the most efficient of all RAL estimators. Recall that our estimator satisfies (1). Thus, if the optimal influence function $\boldsymbol{\varphi}_{\text{eff}}$ is found, then by the central limit theorem, the semiparametric lower bound is given as $E(\boldsymbol{\varphi}_{\text{eff}} \boldsymbol{\varphi}_{\text{eff}}^\top)$. We begin by giving the efficient influence function in Theorem 1. In the following discussion, we may abbreviate the parameter value or random variable, for example, $\pi(\mathbf{Z}; \boldsymbol{\phi}_0) = \pi(\mathbf{Z}) = \pi(\boldsymbol{\phi}_0)$, unless this would lead to ambiguity. Proof of the following theorem is given in Appendix B.

Theorem 1. Let $\mathbf{S}_{\text{eff}} = (\mathbf{S}_1^\top, S_2)^\top$, where $\mathbf{S}_1 = \mathbf{S}_1(R, \mathbf{G}_R(\mathbf{Z}))$ and $S_2 = S_2(R, \mathbf{G}_R(\mathbf{Z}))$ are defined as

$$\mathbf{S}_1(R, \mathbf{G}_R(\mathbf{Z}); \phi) = \left(1 - \frac{R}{\pi(\mathbf{Z}; \phi)}\right) \mathbf{g}^*(\mathbf{X}), \quad (5)$$

$$S_2(R, \mathbf{G}_R(\mathbf{Z}); \phi, \theta) = \frac{R}{\pi(\mathbf{Z}; \phi)}(\theta - U(\mathbf{Z})) + \left(1 - \frac{R}{\pi(\mathbf{Z}; \phi)}\right)(\theta - U^*(\mathbf{X})), \quad (6)$$

$\mathbf{g}^*(\mathbf{x}) = E^*(s_0(\mathbf{Z}; \phi_0) \mid \mathbf{x})$, $U^*(\mathbf{x}) = E^*(U(\mathbf{Z}) \mid \mathbf{x})$, and

$$E^*(\mathbf{g}(\mathbf{Z}) \mid \mathbf{X}) = \frac{E(O(\mathbf{Z})\mathbf{g}(\mathbf{Z}) \mid \mathbf{X})}{E(O(\mathbf{Z}) \mid \mathbf{X})} \quad (7)$$

with $O(\mathbf{z}) = \{1 - \pi(\mathbf{z}; \phi)\}/\pi(\mathbf{z}; \phi)$. Then, the efficient influence function is $\boldsymbol{\varphi}_{\text{eff}} = \mathbf{H}^{-1}\mathbf{S}_{\text{eff}}$, where

$$\mathbf{H} = E(\mathbf{S}_{\text{eff}}^{\otimes 2}) = E\left(\frac{\partial \mathbf{S}_{\text{eff}}(\phi_0, \theta_0)}{\partial (\phi^\top, \theta)^\top}\right)$$

and $\mathbf{B}^{\otimes 2} = \mathbf{B}\mathbf{B}^\top$. Therefore, the semiparametric efficiency bound is given by $\{E(\mathbf{S}_{\text{eff}}^{\otimes 2})\}^{-1}$. In particular, the asymptotic variance of $\hat{\theta}$ is $V = \text{var}\{S_2(\phi_0) - \boldsymbol{\kappa}\mathbf{S}_1(\phi_0, \theta_0)\}$, where $\boldsymbol{\kappa} = E\{(E^*(U \mid \mathbf{X}) - U(\mathbf{Z}))\dot{\pi}(\phi_0)^\top/\pi(\phi_0)\}E\{\mathbf{g}^*(\phi_0)\dot{\pi}(\phi_0)^\top/\pi(\phi_0)\}^{-1}$.

This theorem implies that if we can compute $E^*(\cdot \mid \mathbf{x})$, then equations (5) and (6) give an optimal estimator. This will be a solution of

$$\sum_{i=1}^n \mathbf{S}_{\text{eff},i}(\phi, \theta) = \sum_{i=1}^n (\mathbf{S}_1^\top(R_i, \mathbf{G}_{R_i}(\mathbf{Z}_i); \phi), S_2(R_i, \mathbf{G}_{R_i}(\mathbf{Z}_i); \phi, \theta))^\top = \mathbf{0}, \quad (8)$$

because in the regularity conditions defined in Appendix A, we assume \mathbf{H} is nonsingular; also note that it is well known that multiplying a nonsingular matrix does not affect its asymptotic distribution.

Remark 1. Equation (5) is a special case of the estimator of Chang & Kott (2008). Thus, the optimal \mathbf{g} function in (2) for the Chang & Kott (2008) method is given by $\mathbf{g}^*(\mathbf{X})$ in (5).

Remark 2. Equation (6) has a form similar to that of the doubly robust estimator (Robins et al. 1994) that has been proposed when the response mechanism is missing at random ; unfortunately, our estimator does not have the doubly robustness property.

The equation based on $\mathbf{S}_1(\phi)$ gives an optimal estimator for ϕ , say $\hat{\phi}$. Then, by using $\hat{\phi}$, $S_2(\hat{\phi}, \theta)$ provides an optimal estimator for θ . However, the expectation is unknown and is to be estimated. To perform these computations, we may need to correctly specify the distribution of $Y \mid \mathbf{X}$, which is subjective and unverifiable, as is stated in Section 1. We avoid this problem and propose two adaptive estimators that also attain the lower bound derived in Theorem 1.

4 Adaptive Estimators

We now propose two adaptive estimators for (ϕ_0, θ_0) : (i) with a parametric working model for $f_1(y \mid \mathbf{x})$; (ii) with a nonparametric estimator for $f_1(y \mid \mathbf{x})$.

To discuss the first method, let $f_1(y \mid \mathbf{x})$ be known up to the parameter $\gamma \in \Gamma$, and let $\hat{\gamma}$ be the maximum-likelihood estimator obtained using only observed data, i.e., let it be a solution of

$$\sum_{i=1}^n r_i \log \frac{\partial \log f_1(y_i \mid \mathbf{x}_i; \gamma)}{\partial \gamma} = \mathbf{0}. \quad (9)$$

This can be easily implemented, and its validity can be checked by using an information criterion, such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC). By using the idea similar to that used to derive (4), we can show that, for any function $\mathbf{g}(\mathbf{Z})$,

$$E^*(\mathbf{g}(\mathbf{Z}) \mid \mathbf{x}; \gamma, \phi) = \frac{E_1(\pi^{-1}(\mathbf{Z}; \phi) O(\mathbf{Z}; \phi) \mathbf{g}(\mathbf{Z}) \mid \mathbf{x}; \gamma)}{E_1(\pi^{-1}(\mathbf{Z}; \phi) O(\mathbf{Z}; \phi) \mid \mathbf{x}; \gamma)}. \quad (10)$$

Thus, the expectation can be estimated by using the assumed model $f_1(y \mid \mathbf{x}; \hat{\boldsymbol{\gamma}})$. Although it seems that misspecification of $f_1(y \mid \mathbf{x})$ leads to inconsistency in $(\boldsymbol{\phi}_0, \theta_0)$, to our surprise, the estimator that uses the function on the right-hand side of (10) is consistent, even if it is misspecified. We note that if the model is correctly specified, the estimator attains the lower bound.

Theorem 2. Let $(\hat{\boldsymbol{\phi}}^\top, \hat{\theta})^\top$ be the solution to $\sum_{i=1}^n \mathbf{S}_{\text{eff},i}(\boldsymbol{\phi}, \theta, \hat{\boldsymbol{\gamma}}) = 0$, where $\mathbf{S}_{\text{eff},i}(\boldsymbol{\phi}, \theta, \hat{\boldsymbol{\gamma}})$ is a solution of (8) using the estimation $f_1(y \mid \mathbf{x}; \hat{\boldsymbol{\gamma}})$. Under Conditions (C1)–(C9) given in Appendix A, $(\hat{\boldsymbol{\phi}}^\top, \hat{\theta})^\top$ has consistency and asymptotic normality with variance

$$E \left(\frac{\partial \mathbf{S}_{\text{eff}}^*}{\partial (\boldsymbol{\phi}^\top, \theta)} \right)^{-1} E(\mathbf{S}_{\text{eff}}^{*\otimes 2}) E \left(\frac{\partial \mathbf{S}_{\text{eff}}^*}{\partial (\boldsymbol{\phi}^\top, \theta)} \right)^{-1, \top}, \quad (11)$$

where $\boldsymbol{\gamma}^*$ is the limit of $\hat{\boldsymbol{\gamma}}$, and $\mathbf{S}_{\text{eff}}^* = \mathbf{S}_{\text{eff}}(\boldsymbol{\phi}_0, \theta_0, \boldsymbol{\gamma}^*)$, even if $f_1(y \mid \mathbf{x}; \hat{\boldsymbol{\gamma}})$ is misspecified. Also, the asymptotic variance of $\hat{\theta}$ is given as

$$V^* = \text{var}\{S_2(\boldsymbol{\phi}_0, \theta_0, \boldsymbol{\gamma}^*) - \boldsymbol{\kappa}^* \mathbf{S}_1(\boldsymbol{\phi}_0, \boldsymbol{\gamma}^*)\}, \quad (12)$$

where $\boldsymbol{\kappa}^* = \boldsymbol{\kappa}_1^*(\boldsymbol{\kappa}_2^*)^{-1}$, $\boldsymbol{\kappa}_1^* = E\{(U^*(\boldsymbol{\gamma}^*) - U)\dot{\pi}(\boldsymbol{\phi}_0)^\top / \pi(\boldsymbol{\phi}_0)\}$, and $\boldsymbol{\kappa}_2^* = E\{\mathbf{g}^*(\boldsymbol{\phi}_0, \boldsymbol{\gamma}^*)\dot{\pi}(\boldsymbol{\phi}_0)^\top / \pi(\boldsymbol{\phi}_0)\}$. In addition, if the model is correctly specified, the estimator attains the semiparametric efficiency bound.

Unlike the estimator of Riddles et al. (2016), the parametric model f_1 is irrelevant when considering the consistency and asymptotic normality of our estimator. Therefore, we call f_1 a working model, as in Liang & Zeger (1986).

Numerical computation is needed to calculate the conditional expectation in (10). The expectation-maximization (EM) algorithm considered in Riddles et al. (2016) can be used with a minor modification. We can directly apply their method, once the weights w_{ij}^* defined in (15) of Riddles et al. (2016) are changed to

$$w_{ij}^* = \frac{\pi^{-1}(\mathbf{x}_i, y_j; \boldsymbol{\phi}) O(\mathbf{x}_i, y_j; \boldsymbol{\phi}) f_1(y_j \mid \mathbf{x}_i, \boldsymbol{\gamma}) / C(y_j; \boldsymbol{\gamma})}{\sum_{k: \delta_k=1} \pi^{-1}(\mathbf{x}_i, y_k; \boldsymbol{\phi}) O(\mathbf{x}_i, y_k; \boldsymbol{\phi}) f_1(y_k \mid \mathbf{x}_i, \boldsymbol{\gamma}) / C(y_k; \boldsymbol{\gamma})},$$

where $C(y; \gamma) = \sum_{l: \delta_l=1} f_1(y \mid \mathbf{x}_l; \gamma)$. The weight w_{ij}^* can be called fractional weights in the context of fractional imputation (Kim 2011). By using the above weights, $E^*(\mathbf{g}(\mathbf{Z}) \mid \mathbf{x}_i; \gamma, \phi)$ can be computed by $\sum_{j=1}^n w_{ij}^* \mathbf{g}(\mathbf{z}_j)$.

There exists a class of the response mechanisms and $f_1(y \mid \mathbf{x})$ for which the expectation can be computed analytically when the parameter of interest is the mean of the response variable, i.e., $U(\mathbf{Z}) = Y$. The following proposition would be useful in practice.

Proposition 1. Let $f_1(y \mid \mathbf{x})$ be a function in the exponential family, that is,

$$f_1(y \mid \mathbf{x}; \tau, \psi) = \exp \left(\frac{y\tau - b(\tau)}{\psi} + c(y, \psi) \right), \quad (13)$$

and let its mean be $\lambda\{\mu(\tau \mid \mathbf{x}; \gamma)\} = \gamma^\top \mathbf{x}$, where λ is a link function, and b and c are known functions. Also, let the response mechanism can be written as

$$\pi(\mathbf{x}, y; \phi) = \frac{1}{1 + \exp\{h(\mathbf{x}; \phi_0) + \phi_1 y\}}, \quad (14)$$

where $\phi = (\phi_0^\top, \phi_1)^\top$, and h is a known function. Then, it holds that

$$\mathbf{g}^*(\mathbf{x})^\top = -\frac{(\dot{h}(\mathbf{x}; \phi_0)^\top, \dot{b}(\phi_1\psi + \tau))}{1 + d(\phi_1, \psi, \tau)}$$

and

$$E^*(Y \mid \mathbf{x}) = \frac{\dot{b}(\phi_1\psi + \tau) + d(\phi_1, \psi, \tau)\dot{b}(2\phi_1\psi + \tau)}{1 + d(\phi_1, \psi, \tau)},$$

where $\dot{b}(\tau) = \partial b(\tau)/\partial \tau$ and $d(\phi_1, \psi, \tau) = \exp\{h(\mathbf{x}; \phi_0) + (b(2\phi_1\psi + \tau) - b(\phi_1\psi + \tau))/\psi\}$.

The proof is given in Appendix B. In particular, assume that $Y \mid \mathbf{X}, R = 1$ is normally distributed with mean μ and variance σ^2 , i.e., $\tau = \mu = \gamma^\top \mathbf{x}$, $\psi = \sigma^2$, $b(\tau) = \tau^2/2$, and the response mechanism is $\text{logit}\{\pi(\mathbf{x}, y)\} = h(\mathbf{x}; \phi_0) + \phi_1 y$. Then, it holds that

$$\mathbf{g}^*(\mathbf{x}) = \frac{1}{1 + \exp\{h(\mathbf{x}; \phi_0) + \phi_1(3\phi_1\sigma^2/2 + \gamma^\top \mathbf{x})\}} \begin{pmatrix} \dot{h}(\mathbf{x}; \phi_0) \\ \phi_1\sigma^2 + \gamma^\top \mathbf{x} \end{pmatrix}$$

and

$$E^*(Y \mid \mathbf{x}) = \frac{(\phi_1 \sigma^2 + \boldsymbol{\gamma}^\top \mathbf{x}) + \exp\{h(x; \phi_0) + \phi_1(3\phi_1 \sigma^2/2 + \boldsymbol{\gamma}^\top \mathbf{x})\}(2\phi_1 \sigma^2 + \boldsymbol{\gamma}^\top \mathbf{x})}{1 + \exp\{h(x; \phi_0) + \phi_1(3\phi_1 \sigma^2/2 + \boldsymbol{\gamma}^\top \mathbf{x})\}}.$$

Thus, in this case, we can compute the conditional expectation analytically.

We now discuss the second adaptive estimation method based on the nonparametric estimation for $f_1(y \mid \mathbf{x})$. Generally speaking, directly computing the expectation defined in (7) with the nonparametrically estimated $f_1(y \mid \mathbf{x})$ would make the computation difficult. To avoid this problem, we consider methods of calculating the expectation directly. When x is discrete, such as when x is a dichotomous variable, the expectation can be computed by averaging the data conditioned by $X = x$ and $R = 1$, e.g., for $x = 0, 1$,

$$\hat{E}^*(\mathbf{g}(x, Y) \mid x; \boldsymbol{\phi}) = \frac{\sum_{j \in I_x} r_j \pi^{-1}(x, y_j; \boldsymbol{\phi}) O(x, y_j; \boldsymbol{\phi}) \mathbf{g}(x, y_j)}{\sum_{j \in I_x} r_j \pi^{-1}(x, y_j; \boldsymbol{\phi}) O(x, y_j; \boldsymbol{\phi})} \quad (15)$$

is a consistent estimator of (10), where $I_x = \{j \in \{1, \dots, n\} \mid X_j = x\}$.

A natural extension to the case when \mathbf{X} contains continuous variables can be accomplished with the Nadaraya-Watson estimator. That is, for any function $\mathbf{g}(\mathbf{Z})$,

$$\hat{E}^*(\mathbf{g}(\mathbf{Z}) \mid \mathbf{X}) = \frac{\sum_{j=1}^n r_j K_h(\mathbf{x} - \mathbf{x}_j) r_j \pi^{-1}(\mathbf{x}, y_j; \boldsymbol{\phi}) O(\mathbf{x}, y_j; \boldsymbol{\phi}) \mathbf{g}(\mathbf{x}, y_j)}{\sum_{j=1}^n K_h(\mathbf{x} - \mathbf{x}_j) r_j \pi^{-1}(\mathbf{x}, y_j; \boldsymbol{\phi}) O(\mathbf{x}, y_j; \boldsymbol{\phi})} \quad (16)$$

is consistent under some regularity conditions given in Appendix A. Here, $K_h(\mathbf{x} - \mathbf{w}) = K((\mathbf{x} - \mathbf{w})/h)$, K is a kernel function, and h is the bandwidth. Then, we have the following result for the adaptive estimators obtained with the Nadaraya-Watson estimation.

Theorem 3. Let $(\hat{\boldsymbol{\phi}}^\top, \hat{\theta})^\top$ be a solution of $\sum_{i=1}^n \hat{\mathbf{S}}_{\text{eff},i}(\boldsymbol{\phi}, \theta) = \mathbf{0}$, where $\hat{\mathbf{S}}_{\text{eff},i}(\boldsymbol{\phi}, \theta)$ is a solution of (8) with the estimated conditional expectation defined in (16). Under Conditions (C1)–(C8) and (C10)–(C14) given in Appendix A, $(\hat{\boldsymbol{\phi}}^\top, \hat{\theta})^\top$ satisfies consistency and asymptotic normality, and the estimator attains the semiparametric efficiency bound.

The proposed estimator is attractive because it does not need any assumptions on f_1 , but it would not work well when the dimension of \mathbf{X} is high. This is a common problem in nonparametric estimation.

Remark 3. When Y is also discrete, the fully nonparametric approach proposed in Appendix 2 of Riddles et al. (2016) can be used. For example, if both X and Y are binary, $\pi(x, y) = P(R = 1 \mid x, y)$ takes four possible values. Since the fully nonparametric approach is based on the maximum likelihood, it is the same as the solution of (4) if the estimates of all $\pi(x, y)$ satisfy $0 < \hat{\pi}(x, y) < 1$; otherwise, there are no solutions. However, the estimator of the fully nonparametric approach always has a solution. In that sense, the fully nonparametric approach is better than our proposed method. Our proposed estimator also does not have a solution when it does not satisfy the conditions. Surprisingly, however, the fully nonparametric approach has exactly the same solution as our proposed estimator. The proof is given in Appendix C. Therefore, when X and Y are discrete, the fully nonparametric approach always has a solution, and it gives the best estimator.

Estimation of the variance is also a difficult problem in semiparametric estimation, because it requires some nonparametric estimates other than the parameters of interest. To estimate such functions, a common approach is to use the Horvitz-Thompson type estimator with an estimated response mechanism, but it can be unstable (see Morikawa et al. 2016, §4). However, by using Theorem 1 and Theorem 2, a much more stable variance estimator for V^* can be obtained.

When we consider a parametric working model $f_1(y \mid \mathbf{x}; \boldsymbol{\gamma})$, it follows that

$$\hat{V} = n^{-1} \sum_{i=1}^n \{S_2(R_i, \mathbf{G}_{R_i}(\mathbf{Z}_i); \hat{\theta}, \hat{\phi}, \hat{\gamma}) - \hat{\kappa} \mathbf{S}_1(R_i, \mathbf{G}_{R_i}(\mathbf{Z}_i); \hat{\phi}, \hat{\gamma})\}^{\otimes 2} \quad (17)$$

converges to V^* in probability, as defined in (12), where $\hat{\kappa}$ is a consistent estimator for $\boldsymbol{\kappa}^* = \boldsymbol{\kappa}_1^*(\boldsymbol{\kappa}_2^*)^{-1}$ in the nonparametric estimates, for $\boldsymbol{\kappa}_1^*$ and $\boldsymbol{\kappa}_2^*$ as defined in Theorem 2. To

estimate $\boldsymbol{\kappa}_1^*$, we propose using the same method that we used to compute θ_0 , i.e., let the expectation of $\mathcal{U}(\boldsymbol{\phi}_0, \boldsymbol{\gamma}^*) = (U^*(\boldsymbol{\gamma}^*) - U)\dot{\pi}(\boldsymbol{\phi}_0)^\top / \pi(\boldsymbol{\phi}_0)$ be our target function and solve the following equation:

$$\sum_{i=1}^n \left[\frac{r_i \{\boldsymbol{\kappa}_1^* - \mathcal{U}(\mathbf{z}_i; \boldsymbol{\kappa}_1^*, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\gamma}})\}}{\pi(\mathbf{z}_i; \hat{\boldsymbol{\phi}})} + \left(1 - \frac{r_i}{\pi(\mathbf{z}_i; \hat{\boldsymbol{\phi}})}\right) \{\boldsymbol{\kappa}_1^* - E^*(\mathcal{U}(\mathbf{Z}, \boldsymbol{\kappa}_1^*, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\gamma}}) \mid \mathbf{x}_i; \hat{\boldsymbol{\gamma}})\} \right] = \mathbf{0}.$$

This is the best estimator for $(\boldsymbol{\phi}, \boldsymbol{\kappa}_1^*)$ in terms of the asymptotic variance, because the second term in \mathcal{U} is known, and so we can apply Theorem 2. The best estimator for $\boldsymbol{\kappa}_2^*$ can be obtained in the same way. When we use the method of Theorem 3 to estimate θ , a similar method can be used to estimate the variance.

5 Simulation Study

In order to evaluate the performance of our estimators with a finite sample, we conducted a Monte Carlo simulation with two scenarios, Scenario s for $s = 1, 2$. In each scenario, we used two covariate variables $\mathbf{X} = (X_1, X_2)^\top$, set the response mechanism to $\text{logit}\{\pi(x_1, y)\} = \phi_0 + \phi_1 x_1 + \phi_2 y$, and generated the response variable from $Y \mid (x_1, x_2, r = 1) \sim N(\mu_s(\mathbf{x}), 1)$.

In Scenario 1, both X_1 and X_2 are binary with generation probability 0.5 and $\mu_1(\mathbf{x}) = 0.5I_{00} - 0.5I_{01} - 0.5I_{10} + I_{11}$, respectively, where $I_{ij} = I(X_1 = i, X_2 = j)$, and I is an indicator function. The parameter of the response model is $(\phi_0, \phi_1, \phi_2) = (-1, -0.5, 0.75)$. In Scenario 2, X_1 is generated from the same distribution as in Scenario 1, but X_2 is generated from $U(-1, 1)$, where $U(a, b)$ is a uniform distribution with minimum value a and maximum value b . The mean structure is $\log\{\mu_2(\mathbf{x})\} = -1.7 - 0.4x_1 + 0.5x_2$ and $(\phi_0, \phi_1, \phi_2) = (-1.7, -0.4, 0.5)$. We note that in both scenarios, the response rate is about 70%, and X_2 is a nonresponse instrumental variable (Wang et al. 2014). We are interested in estimating response model parameter and the population parameter $\theta = E(Y)$. For

response models, only ϕ_2 is reported.

Data (R, \mathbf{X}, Y) are generated as follows: (i) \mathbf{X} is generated by the assumed model; (ii) R is generated by $P(R = 1 | \mathbf{x})$, which is equal to $E_1\{\pi^{-1}(x_1, Y) | \mathbf{x}\}^{-1}$ (thus, it can be computed by the assumed response model and f_1), because by Bayes' formula,

$$\begin{aligned} E_1\{\pi^{-1}(x_1, Y) | \mathbf{x}\} &= \int \pi^{-1}(x_1, y) f_1(y | \mathbf{x}) dy \\ &= \int \pi^{-1}(x_1, y) \frac{\pi(x_1, y) f(y | \mathbf{x})}{P(R = 1 | \mathbf{x})} dy = \{P(R = 1 | \mathbf{x})\}^{-1}. \end{aligned}$$

When f_1 belongs to the exponential family defined in (13), and the response model is the logistic model defined in (14), it follows that

$$P(R = 1 | \mathbf{x}) = \frac{1}{1 + \exp[h(\mathbf{x}) + \{b(\phi\psi + \tau) - b(\tau)\}/\psi]};$$

(iii) when $r = 1$, Y is generated by $f_1(y | \mathbf{x})$, and $r = 0$, Y is from

$$\begin{aligned} f_0(y | \mathbf{x}) &= \frac{O(x_1, y)}{E_1\{O(x_1, Y) | \mathbf{x}\}} f_1(y | \mathbf{x}) \\ &= \exp \left\{ \frac{y(\phi\psi + \tau) - b(\phi\psi + \tau)}{\psi} + c(y, \psi) \right\}, \end{aligned}$$

which implies $f_0(y | \mathbf{x})$ also belongs to the exponential family, and thus Y can be easily generated.

From each sample, we computed four estimators, as follows:

- [1] MAR: A naive estimator based on the assumption that the missing data are missing-at-random:

$$\sum_{i=1}^n \delta_i(\theta - y_i) / \hat{\pi}_i = 0, \tag{18}$$

where $\hat{\pi}_i$ is an estimated response model under this assumption, that is, $\hat{\pi} = \{1 + \exp(-\hat{\phi}_0 - \hat{\phi}_1 x_1 - \hat{\phi}_2 x_2)\}^{-1}$, where $(\hat{\phi}_0, \hat{\phi}_1, \hat{\phi}_2)$ is the maximum-likelihood estimator.

- [2] CK: The estimator of Chang & Kott (2008). We use the estimating equation (2), setting \mathbf{g} as $(1, x_1, x_2)$; θ is estimated using the ϕ estimated by (18) .
- [3] RKI: The estimator of Riddles et al. (2016). In Scenario 1, there is no need for a parametric model f_1 , and a nonparametric method similar to (15) is used; however, in Scenario 2, it is necessary to assume a misspecified parametric model f_1 that is normally distributed with a mean structure of $\gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_2^2 + \gamma_4 x_1 x_2$. The other model is chosen by the AIC.
- [4] New: Our proposed estimator. In Scenario 2, we also use the same misspecified model as is used for the RKI, say [Nm], and we also assume a correctly specified model [Nt] and the nonparametric estimator proposed in (15) [Nnp].

Monte Carlo samples of size $n = 500$ and $2,000$ were independently generated $2,000$ times. We used the correct model for the response mechanism, except with MAR.

Remark 4. Qin et al. (2002) also derived an estimator that is effective for nonignorable nonresponse data. However, their estimator is exactly the same as that of Chang & Kott (2008) when $q = d + 1$ and the response model is logistic, where q and d are the dimensions of ϕ and x , respectively. The proof is shown in Appendix C. For this reason, we omitted the results using the estimator of Qin et al. (2002).

Tables 1 and 2 show the Monte Carlo simulation results for Scenarios 1 and 2, respectively. In Tables 1 and 2, the bias and standard errors (S.E.) of the point estimators are presented for each case. In the CK method, we encountered some numerical problems in Scenario 1, and there was no solution because the estimate of ϕ did not converge. This was true even for a large sample. The following is a summary of the simulation results shown in Tables 1 and 2:

Table 1: Monte Carlo bias and S.E. of the four estimators, based on 2,000 Monte Carlo samples for Scenario 1. All values $\times 1000$.

Scenario	Size	Parameter	MAR	CK	RKI	New
1	500	ϕ_2	Bias	-278	7	3
			S.E.	1254	198	187
		θ	Bias	-314	-140	-3
			S.E.	64	391	88
			#NA	0	785	0
	2000	ϕ_2	Bias	-254	4	3
			S.E.	1147	89	89
		θ	Bias	-314	-112	0
			S.E.	31	356	42
			#NA	0	722	0

#NA: total number of not applicable (due to numerical problems).

Table 2: Monte Carlo bias and S.E. of the six estimators, based on 2,000 Monte Carlo samples for Scenario 2. Nm, Nt, and Nnp are our proposed estimators. All values $\times 1000$.

Scenario	Size	Parameter	MAR	CK	RKI	New		
						Nm	Nt	Nnp
2	500	ϕ_2	Bias	15	31	9	7	0
			S.E.	100	93	91	87	87
		θ	Bias	-300	-10	9	-7	-23
			S.E.	94	104	125	99	100
	2000	ϕ_2	Bias	4	24	3	3	-1
			S.E.	44	42	43	41	41
		θ	Bias	-297	0	0	1	-6
			S.E.	47	53	63	49	49

Table 3: Coverage probability based on 2,000 Monte Carlo samples. New is the estimator in Scenario 1 and the others are in Scenario 2. All values $\times 100$.

Size	New	Nm	Nt	Nnp
500	94.25	93.40	94.20	87.20
2000	95.05	94.65	95.35	92.35

- [1] In all scenarios, the naive estimator using the MAR assumption is significantly biased, since this assumption does not hold.
- [2] The CK method does not work well in Scenario 1, because both X_1 and X_2 are binary, and constraints only on the binary covariates are not sufficient for estimating ϕ . However, its performance is good in Scenario 2.
- [3] The RKI method and our estimators perform well in Scenario 1. However, in Scenario 2, the estimates for θ are somewhat positively biased in RKI due to the misspecification of f_1 .
- [4] In all scenarios, our proposed estimators perform better than any other methods. We note that in Scenario 2, the estimators of Nt and Nnp have the smallest variance, but the bias of the nonparametric estimator of θ is a little larger than that of the other estimators; however, it decreases as the sample size increases. The Nm model is consistent even if the model on f_1 is misspecified; however, it is less efficient than the other proposed estimators.

Table 3 shows the estimated coverage probability of our proposed estimators, for significance level $\alpha = 0.05$. We used (17) to estimate the variance of our estimators. Our proposed variance estimator works well, except for nonparametric methods, because it is biased when the sample size is not very large. However, as the sample size increases, the coverage probability approaches to 95%.

6 Analysis of Real Data

In this section, our proposed estimators are applied to the Korea Labor and Income Panel Survey (KLIPS), which has been analyzed multiple times (Kim & Yu 2011, Wang et al.

2014, Shao & Wang 2016). The data represent $n = 2,506$ Korean residents; the response variable y is income (10^6 Korean Won) in year 2008. There are four covariates: x_1 : age; x_2 : income in the previous year (2007); x_3 : gender; and x_4 : education level. In our analysis, we considered three data points to be outliers, and they were excluded in our analysis.

Since the data are completely observed, we made eight artificial incomplete datasets by assuming the eight response models considered in Kim & Yu (2011): M1 (linear ignorable): $\text{logit}(\pi) = 4.93 - x_1$; M2 (linear nonignorable): $\text{logit}(\pi) = 1.61 - 0.5x_1 + 0.75y$; M3 (nonlinear nonignorable, quadratic in x_1): $\text{logit}(\pi) = 3.7 - 0.5x_1 - 0.1x_1^2 + 0.6y$; M4 (jump nonignorable): $\pi = 0.5I(y \leq 1.6) + I(y > 1.6)$; M5 (nonlinear nonignorable, quadratic in y): $\text{logit}(\pi) = -0.22 + 0.1x_1 + 0.1y + 0.2y^2$; M6 (probit nonignorable): $\pi = \Phi(0.71 - 0.25x_1 + 0.5y)$, where Φ is the cumulative distribution function of the normal distribution; M7 (complementary log-log nonignorable): $\pi = 1 - \exp\{-\exp(-1.8 + 0.2x_1 + 0.8y)\}$; and M8 (nonlinear nonignorable, interaction): $\text{logit}(\pi) = -0.2 - 0.1x_1 + 0.2y + 0.2x_1y$. For all data sets, the response rate is about 70%.

We assumed the response model $\text{logit}\{\pi(x_1, y; \phi)\} = \phi_0 + \phi_1x_1 + \phi_2x_1^2 + \phi_3y + \phi_4y^2$, which implies, for models M4, M6, M7, and M8, that our response model misspecified the true model. We assumed that the use of a working model f_1 would be better, since there are four covariate variables and the sample size is not very large. Thus, f_1 is specified as the normal distribution with mean structure $\mu(x) = \gamma_0 + \sum_{ij} \gamma_{ij}x_ix_j$, and one model is chosen by the AIC. The mean income in 2008 was 1.648, as calculated using the complete data.

The resulting estimates and estimated confidence intervals are shown in Table 4. It can be seen that our estimators and estimated confidence interval work well for all datasets, including when the model is misspecified.

Table 4: Estimates, bias, and confidence interval with significance level $\alpha = 0.05, 0.01$ are reported for datasets M1–M8, based on real data. Bias values $\times 1000$.

Missing	Bias	CI(0.05)	CI(0.01)
M1	-5	(1.605, 1.682)	(1.592, 1.694)
M2	-8	(1.603, 1.677)	(1.592, 1.688)
M3	2	(1.600, 1.700)	(1.584, 1.716)
M4	-3	(1.605, 1.685)	(1.592, 1.698)
M5	5	(1.614, 1.693)	(1.601, 1.705)
M6	-7	(1.604, 1.679)	(1.593, 1.690)
M7	-3	(1.609, 1.682)	(1.597, 1.694)
M8	-6	(1.606, 1.679)	(1.594, 1.691)

7 Discussion

We have derived a semiparametric efficiency bound for $(\phi^\top, \theta)^\top$ under a nonignorable nonresponse. We also proposed two types of adaptive semiparametric estimators that each attain the lower bound. One uses a working model on $f_1(y \mid \mathbf{x}) = f(y \mid \mathbf{x}, r = 1)$. If the model is correctly specified, our proposed estimator attains the bound, and it has consistency and asymptotic normality even if it is misspecified. Validity of the working model can be checked by using an existing information criteria, such as the AIC or BIC, because the model is based on observed data. The second estimator utilizes the Nadaraya-Watson estimator, which is nonparametric; it always attains the bound. Other nonparametric estimators, such as the nonparametric Bayes, and techniques of machine learning can be used instead of the Nadaraya-Watson estimator. The second estimator is

better than the first one because it attains the lower bound, even though it does not need the specification of f_1 ; however, it does not work well when the dimension of \mathbf{X} is high. We have also clarified the relationships between the our estimator and previous ones, including that of Qin et al. (2002), Chang & Kott (2008), and Riddles et al. (2016).

The proposed method is based on the correct specification of the response model. There may be various candidates for the response model, and thus the appropriate information criteria for choosing the response model will be a topic of future research. Instead of specifying a single response model, one can consider multiple response models, and obtain consistency when one of the specified response models is correct. This multiple robustness property has been investigated in the ignorable nonresponse setup (Han 2014). Extension of multiple robustness to the nonignorable nonresponse case will also be a topic of our future research.

Appendix A

Following conditions are needed to guarantee Theorem 1–3:

- (C1). There exists a nonresponse instrumental variable \mathbf{X}_2 , i.e., $\mathbf{X} = (\mathbf{X}_1^\top, \mathbf{X}_2^\top)^\top$, and \mathbf{X}_2 is independent of R given \mathbf{X}_1 and Y .
- (C2). The response probability $\pi(\mathbf{x}_{1i}, y_i)$ is bounded below. That is, $\pi(\mathbf{x}_{1i}, y_i) > K$ for some $K > 0$ for all $i = 1, \dots, n$, uniformly in n .
- (C3). $\mathbf{W}_i = (\mathbf{X}_i, Y_i, R_i)$ are independently and identically distributed.
- (C4). Φ and Θ are compact.

(C5). $\mathbf{S}_{\text{eff}}(\boldsymbol{\phi}, \theta)$ defined in (5), (6), and (8) is continuously differentiable at each $(\boldsymbol{\phi}, \theta) \in \Phi \times \Theta$ with probability one, and there exists $d(\mathbf{W})$ such that $\|\mathbf{S}_{\text{eff}}(\boldsymbol{\phi}, \theta)\| \leq d(\mathbf{W})$ for all $(\boldsymbol{\phi}, \theta) \in \Phi \times \Theta$ and $E\{d(\mathbf{W})\} < \infty$.

(C6). $E\{\mathbf{S}_{\text{eff}}(\boldsymbol{\phi}, \theta)\} = \mathbf{0}$ has a unique solution $(\boldsymbol{\phi}_0, \theta_0) \in \Phi \times \Theta$.

(C7). $\partial \mathbf{S}_{\text{eff}}(\boldsymbol{\phi}, \theta) / \partial(\boldsymbol{\phi}^\top, \theta)$ is continuous at $(\boldsymbol{\phi}_0, \theta_0)$ with probability one, and there is a neighborhood $\Phi_{\mathcal{N}} \times \Theta_{\mathcal{N}}$ of $(\boldsymbol{\phi}_0, \theta_0)$ such that

$$\|E\{\sup_{(\boldsymbol{\phi}, \theta) \in \Phi_{\mathcal{N}} \times \Theta_{\mathcal{N}}} \partial \mathbf{S}_{\text{eff}}(\boldsymbol{\phi}, \theta) / \partial(\boldsymbol{\phi}^\top, \theta)\}\| < \infty.$$

(C8). $E\{\partial \mathbf{S}_{\text{eff}}(\boldsymbol{\phi}, \theta) / \partial(\boldsymbol{\phi}^\top, \theta)\}$ is nonsingular at $(\boldsymbol{\phi}_0, \theta_0)$.

(C9). Γ is compact, $\mathbf{S}_\gamma(\boldsymbol{\gamma}) = \partial \log f_1(y \mid \mathbf{x}; \boldsymbol{\gamma}) / \partial \boldsymbol{\gamma}$ is continuously differentiable at $\boldsymbol{\gamma} \in \Gamma$ with probability one, there exists $e(\mathbf{W})$ such that $\|\mathbf{S}_\gamma(\boldsymbol{\gamma})\| \leq e(\mathbf{W})$ for all $\boldsymbol{\gamma} \in \Gamma$ and $E\{e(\mathbf{W})\} < \infty$, $E\{\mathbf{S}_\gamma(\boldsymbol{\gamma})\} = \mathbf{0}$ has a unique solution $\boldsymbol{\gamma}_0 \in \Gamma$, $\partial \mathbf{S}_\gamma(\boldsymbol{\gamma}) / \partial \boldsymbol{\gamma}^\top$ is continuous at $\boldsymbol{\gamma}_0$ with probability one, and there is a neighborhood $\Gamma_{\mathcal{N}}$ of $\boldsymbol{\gamma}_0$ such that $\|E\{\sup_{\boldsymbol{\gamma} \in \Gamma_{\mathcal{N}}} \partial \mathbf{S}_\gamma(\boldsymbol{\gamma}) / \partial \boldsymbol{\gamma}^\top\}\| < \infty$.

(C10). Let \mathfrak{X} be a compact set that is contained in the support of \mathbf{x} , let $f_1(\mathbf{x}) > 0$, and let $E_1\{\pi(\mathbf{x}_1, Y; \boldsymbol{\phi}_0) \mid \mathbf{x}\} > 0$ for all $\mathbf{x} \in \mathfrak{X}$.

(C11). The kernel $K(\mathbf{u})$ has bounded derivatives of order k , satisfies $\int K(\mathbf{u}) d\mathbf{u} = 1$, has zero moments of order $\leq m - 1$, and has a nonzero m -th order moment.

(C12). For all y , $\pi(\cdot, y; \boldsymbol{\phi}_0)$, $\dot{\pi}(\cdot, y; \boldsymbol{\phi}_0)$, and $U(\cdot, y; \theta_0)$ are differentiable to order k and are bounded on an open set containing \mathfrak{X} .

(C13). Let $a_1(\mathbf{z}) = 1$, $a_2(\mathbf{z}) = \mathbf{s}_0(\mathbf{z}; \boldsymbol{\phi}_0)$, and $a_3(\mathbf{z}) = U(\mathbf{z})$. Then, there exists $v \geq 4$ such that $E_1\{|\pi^{-1}(\mathbf{Z}; \boldsymbol{\phi}_0)O(\mathbf{Z}; \boldsymbol{\phi}_0)a_i(\mathbf{Z})|^v\}$ and $E_1\{\|\pi^{-1}(\mathbf{Z}; \boldsymbol{\phi}_0)O(\mathbf{Z}; \boldsymbol{\phi}_0)a_i(\mathbf{Z})\|^v \mid \mathbf{x}\}f_1(\mathbf{x})$ are bounded for all $\mathbf{x} \in \mathfrak{X}$.

(C14). As $h \rightarrow 0$, $n^{1-(2/v)}h^d/\ln n \rightarrow \infty$, $n^{1/2}h^{d+2k}/\ln n \rightarrow \infty$, and $n^{1/2}h^{2m} \rightarrow 0$.

Appendix B

In this section, we provide a proof for Theorems 1-3; for simplicity, we will assume $U(\mathbf{Z}) = Y$. In order to discuss semiparametric theory, we specify the joint distribution $\mathbf{Z} = (\mathbf{X}^\top, Y)^\top$ by $f(\mathbf{z}; \boldsymbol{\eta})$, where $\boldsymbol{\eta}$ is an infinite-dimensional nuisance parameter, and $\boldsymbol{\eta}_0$ is the true value. By “full model” we refer to the class of models in which the data are completely observed, and by “obs model” we refer to those in which some Y are missing; that is, a full model consists of functions $\mathbf{h}(\mathbf{Z})$ and an obs model consists of $\mathbf{h}(R, \mathbf{G}_R(\mathbf{Z}))$. Furthermore, for each full and obs model, let us denote the nuisance tangent space by Λ^F and Λ , respectively, and its orthogonal complement by $\Lambda^{F\perp}$ and Λ^\perp , respectively. Let S_ϕ be the score function with respect to $\boldsymbol{\phi}$. Consider a Hilbert space $\mathcal{H} = \{\mathbf{h}^{(q+1)\times 1} \mid E(\mathbf{h}) = \mathbf{0}; \|\mathbf{h}\| < \infty\}$ with inner product $\langle \mathbf{h}_1, \mathbf{h}_2 \rangle = E(\mathbf{h}_1^\top \mathbf{h}_2)$, where the expectation is taken under the true model. See Bickel et al. (1998) and Tsiatis (2006) for more details. When U comprises other functions, the proof is almost the same.

In order to derive the efficient score for $(\boldsymbol{\phi}, \theta)$, we introduce a proposition proved in Rotnitzky & Robins (1997), as follows. Let \mathbf{B} and \mathbf{D} be functions of $(R, \mathbf{G}_R(\mathbf{Z}))$, and let \mathbf{B}^* and \mathbf{D}^* be functions of \mathbf{Z} . Also, let us define following three linear operators: $g(\mathbf{B}^*) = E(\mathbf{B}^* \mid R, \mathbf{G}_R(\mathbf{Z}))$, $m(\mathbf{B}^*) = E\{g(\mathbf{B}^*) \mid \mathbf{Z}\}$, and $u(\mathbf{B}^*) = R\mathbf{B}^*/\pi(\mathbf{Z})$. Then, the efficient score for $(\boldsymbol{\phi}, \theta)$ can be derived by the following Lemma. See Proposition A1 in Rotnitzky & Robins (1997) for the proof.

Lemma B1. The efficient score for $(\boldsymbol{\phi}, \theta)$ can be written as

$$\mathbf{S}_{\text{eff}} = \mathbf{u}(\mathbf{D}_{\text{eff}}^*) - \Pi[\mathbf{u}(\mathbf{D}_{\text{eff}}^*) \mid \Lambda_2] + \mathbf{A}_{2,\text{eff}} = \mathbf{g}\{\mathbf{m}^{-1}(\mathbf{D}_{\text{eff}}^*)\} + \mathbf{A}_{2,\text{eff}}, \quad (\text{A.1})$$

where $\Pi[h \mid \Lambda_2]$ is the projection of h onto Λ_2 , $\Lambda_2 = [\mathbf{h}(R, \mathbf{G}_R(\mathbf{Z})) : E(\mathbf{h}(R, \mathbf{G}_R(\mathbf{Z})) \mid \mathbf{Z}) = \mathbf{0}]$, and $\mathbf{D}_{\text{eff}}^*$ is a unique solution of

$$\Pi[\mathbf{m}^{-1}(\mathbf{D}^*) \mid \Lambda^{F\perp}] = (\mathbf{Q}, S_{\text{eff},\theta}^{F\perp}), \quad (\text{A.2})$$

where $\mathbf{Q} = \Pi[\mathbf{m}^{-1}[E\{\mathbf{g}(\mathbf{S}_\phi^F) \mid \mathbf{L}\}] \mid \Lambda^{F\perp}]$, $\mathbf{A}_{2,\text{eff}} = (\Pi[\mathbf{S}_\phi \mid \Lambda_2]^\top, \mathbf{0})^\top = (\mathbf{g}(\mathbf{S}_\phi^F) - \mathbf{g}[\mathbf{m}^{-1}[E\{\mathbf{g}(\mathbf{S}_\phi^F) \mid \mathbf{L}\}]]^\top, \mathbf{0})^\top$, and $S_{\text{eff},\theta}^{F\perp}$ is the efficient score function of θ in the full model.

This Lemma implies that the efficient score can be represented by (A.1) with $\mathbf{D}_{\text{eff}}^*$ satisfying condition (A.2). Thus, in the nonignorable nonresponse case, $\Lambda^{F\perp}$ is to be calculated, and it can be done in a way similar to that shown in Section 4.5 of Tsiatis (2006).

Lemma B2. The nuisance tangent space Λ^F and its orthogonal complement $\Lambda^{F\perp}$ in the full model are written as follows:

$$\Lambda^F = [\mathbf{h}(\mathbf{Z}) \in \mathcal{H} \text{ such that } E\{Y\mathbf{h}(\mathbf{Z})\} = \mathbf{0}], \quad (\text{A.3})$$

$$\Lambda^{F\perp} = [\mathbf{k}(Y - \theta_0), \text{ where } \mathbf{k} \in \mathbb{R}^{q+1}]. \quad (\text{A.4})$$

Finally, we give an explicit formula to calculate the projection onto Λ_2 .

Lemma B3. For $\mathbf{h}(R, \mathbf{G}_R(\mathbf{Z})) = R\mathbf{h}_1(\mathbf{Z}) + (1 - R)\mathbf{h}_2(\mathbf{X})$, it holds that

$$\Pi(\mathbf{h} \mid \Lambda_2) = \left(1 - \frac{R}{\pi(\mathbf{Z})}\right) E^*(\mathbf{h}_2(\mathbf{X}) - \mathbf{h}_1(\mathbf{Z}) \mid \mathbf{X}). \quad (\text{A.5})$$

Proof of Lemma B3. Obviously, the right-hand side of (A.5) belongs to Λ_2 . Thus, it remains to be checked that for any \mathbf{g} ,

$$\left\langle \mathbf{h} - \left(1 - \frac{R}{\pi(\mathbf{Z})}\right) E^*\{\mathbf{h}_2(\mathbf{X}) - \mathbf{h}_1(\mathbf{Z}) \mid \mathbf{X}\}, \left(1 - \frac{R}{\pi(\mathbf{Z})}\right) \mathbf{g}(\mathbf{X}) \right\rangle = 0,$$

which can be proved easily. \square

We now give proofs of Theorem 1–3, and Proposition 1.

Proof of Theorem 1. Note that $S_{\text{eff},\theta}^{F\perp} = Y - \theta_0$ by Lemma B2, since there exists only one influence function, and it is the efficient one under the assumption that θ does not require any assumptions on the distribution of \mathbf{Z} (see Tsiatis 2006, Chap. 5). By the projection theorem, there exists a unique $\mathbf{k} = (k_1, \mathbf{k}_2^\top)^\top$, where $k_1 (\neq 0) \in \mathbb{R}$ and $\mathbf{k}_2 \in \mathbb{R}^q$, such that $\mathbf{D}_{\text{eff}}^* = \mathbf{k}(Y - \theta_0)$.

Then, we calculate $\mathbf{A}_{2,\text{eff}}$. The score function of ϕ is

$$\mathbf{S}_\phi = \mathbf{g}(\mathbf{S}_\phi^F) = R \frac{\dot{\pi}(\phi)}{\pi(\phi)} - (1 - R)E_0 \left(\frac{\dot{\pi}(\phi)}{1 - \pi(\phi)} \mid \mathbf{X} \right).$$

By Lemma B3, $\Pi(\mathbf{S}_\phi \mid \Lambda_2) = -(1 - R/\pi(\mathbf{Z}))g^*(\mathbf{X})$. Thus, $\mathbf{A}_{2,\text{eff}} = (0, -(1 - R/\pi(\mathbf{Z}))g^*(\mathbf{X}))$.

Again, by using Lemma B3, it follows that $\Pi[\mathbf{u}(\mathbf{D}_{\text{eff}}^*) \mid \Lambda_2] = -(1 - R/\pi(\mathbf{Z}))E^*(Y - \theta_0 \mid \mathbf{X})$,

by which (A.1) becomes

$$\mathbf{S}_1 = \mathbf{k}_2 \left\{ \frac{R(Y - \theta_0)}{\pi(\phi_0)} + \left(1 - \frac{R}{\pi(\phi_0)} \right) E^*(Y - \theta_0 \mid \mathbf{X}) \right\} - \left(1 - \frac{R}{\pi(\mathbf{Z})} \right) g^*(\mathbf{X}) \quad (\text{A.6})$$

and

$$S_2 = k_1 \left\{ \frac{R(Y - \theta_0)}{\pi(\phi_0)} + \left(1 - \frac{R}{\pi(\phi_0)} \right) E^*(Y - \theta_0 \mid \mathbf{X}) \right\}. \quad (\text{A.7})$$

This $\mathbf{S}_{\text{eff}} = (\mathbf{S}_1, S_2^\top)$ can be transformed into $\tilde{\mathbf{S}}_{\text{eff}} = (\tilde{\mathbf{S}}_1, \tilde{S}_2^\top) = \mathbf{A}\mathbf{S}_{\text{eff}}$,

$$\begin{aligned} \tilde{\mathbf{S}}_1 &= \left(1 - \frac{R}{\pi(\phi_0)} \right) \mathbf{g}^*(\mathbf{X}), \\ \tilde{S}_2 &= \frac{R(Y - \theta_0)}{\pi(\phi_0)} + \left(1 - \frac{R}{\pi(\phi_0)} \right) E^*(Y - \theta_0 \mid \mathbf{X}) \end{aligned}$$

with a nonsingular matrix \mathbf{A} ,

$$\mathbf{A} = \begin{bmatrix} -\mathbf{I}_q & -\mathbf{k}_2^\top/k_1 \\ \mathbf{0}^\top & k_1^{-1} \end{bmatrix},$$

where \mathbf{I}_q is a q -dimensional identity matrix. The score function multiplied by a nonsingular constant matrix does not have an influence on the asymptotic distribution. Thus, we have the desired efficient score. \square

Proof of Theorem 2. Let $\boldsymbol{\xi} = (\boldsymbol{\phi}^\top, \theta)^\top$. Recall that each $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\xi}}$ is a solution of $\sum_{i=1}^n \partial \log f_1(y_i \mid \mathbf{x}_i; \boldsymbol{\gamma}) / \partial \boldsymbol{\gamma} = \sum_{i=1}^n \mathbf{S}_{\boldsymbol{\gamma}i}(\boldsymbol{\gamma}) = \mathbf{0}$ and (8), respectively. By using standard asymptotic theory,

$$\begin{bmatrix} \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0 \\ \hat{\boldsymbol{\xi}} - \boldsymbol{\xi}_0 \end{bmatrix} = -\mathbf{I}^{-1} n^{-1} \sum_{i=1}^n \begin{bmatrix} \mathbf{S}_{\boldsymbol{\gamma}i}(\boldsymbol{\gamma}_0) \\ \mathbf{S}_{\text{eff},i}(\boldsymbol{\gamma}_0, \boldsymbol{\xi}_0) \end{bmatrix},$$

where

$$\mathbf{I} = E \begin{bmatrix} \partial \mathbf{S}_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}_0) / \boldsymbol{\gamma}^\top & \partial \mathbf{S}_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}_0) / \boldsymbol{\xi}^\top \\ \partial \mathbf{S}_{\text{eff}}(\boldsymbol{\gamma}_0, \boldsymbol{\xi}_0) / \boldsymbol{\gamma}^\top & \partial \mathbf{S}_{\text{eff}}(\boldsymbol{\gamma}_0, \boldsymbol{\xi}_0) / \boldsymbol{\xi}^\top \end{bmatrix} = E \begin{bmatrix} \partial \mathbf{S}_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}_0) / \boldsymbol{\gamma}^\top & \mathbf{O} \\ \partial \mathbf{S}_{\text{eff}}(\boldsymbol{\gamma}_0, \boldsymbol{\xi}_0) / \boldsymbol{\gamma}^\top & \partial \mathbf{S}_{\text{eff}}(\boldsymbol{\gamma}_0, \boldsymbol{\xi}_0) / \boldsymbol{\xi}^\top \end{bmatrix}.$$

Let the (i, j) block of \mathbf{I} be \mathbf{I}_{ij} . Then,

$$\mathbf{I}^{-1} = \begin{bmatrix} \mathbf{I}_{11}^{-1} & \mathbf{O} \\ -\mathbf{I}_{22}^{-1} \mathbf{I}_{21} \mathbf{I}_2^{-1} & \mathbf{I}_{22}^{-1} \end{bmatrix}.$$

Here, it follows that $\mathbf{I}_{21} = \mathbf{O}$ from

$$E \left\{ \left(1 - \frac{R}{\pi(\boldsymbol{\phi}_0)} \right) \frac{\partial \mathbf{g}^*(\boldsymbol{\gamma}_0, \boldsymbol{\phi}_0)}{\partial \boldsymbol{\gamma}^\top} \right\} = \mathbf{O}$$

and

$$E \left\{ \left(1 - \frac{R}{\pi(\boldsymbol{\phi}_0)} \right) \frac{\partial U^*(\boldsymbol{\gamma}_0, \boldsymbol{\phi}_0)}{\partial \boldsymbol{\gamma}^\top} \right\} = \mathbf{0}^\top.$$

Therefore, we have,

$$\mathbf{I}^{-1} = \begin{bmatrix} \mathbf{I}_{11}^{-1} & \mathbf{O} \\ \mathbf{O} & \mathbf{I}_{22}^{-1} \end{bmatrix}.$$

\square

Proof of Proposition 1. It follows from the definition of \mathbf{g}^* and f_1 that

$$\begin{aligned}\mathbf{g}^{*\top}(\mathbf{x}) &= \frac{E_1\{O(\boldsymbol{\phi}; \mathbf{x}, Y)(\dot{\mathbf{h}}(\mathbf{x}; \boldsymbol{\phi}_0)^\top, Y) \mid \mathbf{x}\}}{E_1\{\pi^{-1}(\boldsymbol{\phi}; \mathbf{x}, Y)O(\boldsymbol{\phi}; \mathbf{x}, Y) \mid \mathbf{x}\}} \\ &= \frac{(\dot{\mathbf{h}}(\mathbf{x}; \boldsymbol{\phi}_0)^\top E_1\{\exp(\phi_1 Y) \mid \mathbf{x}\}, E_1\{Y \exp(\phi_1 Y) \mid \mathbf{x}\})}{E_1[\{1 + \exp(h(\mathbf{x}; \boldsymbol{\phi}_0) + \phi_1 Y)\} \exp(\phi_1 Y) \mid \mathbf{x}]},\end{aligned}\tag{A.8}$$

where $\dot{\mathbf{h}}(\boldsymbol{\phi}_0) = \partial \mathbf{h}(\boldsymbol{\phi}_0)/\partial \boldsymbol{\phi}_0$. Thus, $E_1\{\exp(\phi_1 Y) \mid \mathbf{x}\}$ and $E_1\{Y \exp(\phi_1 Y) \mid \mathbf{x}\}$ are to be computed. Since $E_1\{\exp(\phi_1 Y) \mid \mathbf{x}\}$ is the generating moment function of $Y \mid \mathbf{x}$, it holds that

$$E_1(\exp(\phi_1 Y) \mid \mathbf{x}) = \exp\left(\frac{b(\phi_1 \psi + \theta) - b(\theta)}{\psi}\right)$$

and

$$E_1(Y \exp(\phi_1 Y) \mid \mathbf{x}) = \dot{b}(\phi_1 \psi + \theta) \exp\left(\frac{b(\phi_1 \psi + \theta) - b(\theta)}{\psi}\right).$$

Substituting these results into (A.8), we obtain the proof. Computation of $E^*(Y \mid \mathbf{x})$ can be done in a similar way. \square

Proof of Theorem 3. Let $f_1(\mathbf{x})$ be a conditional distribution of X given $R = 1$, and let $\hat{g}^*(\mathbf{x}) = \hat{D}/\hat{C}$, $D = E_1\{\pi^{-1}(\mathbf{Z}; \boldsymbol{\phi})O(\mathbf{Z}; \boldsymbol{\phi}) \mid \mathbf{x}\}f_1(\mathbf{x})P(R = 1)$, and $C = E_1\{\pi^{-1}(\mathbf{Z}; \boldsymbol{\phi})O(\mathbf{Z}; \boldsymbol{\phi})\mathbf{s}_0(\mathbf{Z}; \boldsymbol{\phi}) \mid \mathbf{x}\}f_1(\mathbf{x})P(R = 1)$. From the same arguments that were used to prove Lemma 1 in Morikawa et al. (2016), it can be shown that

$$\frac{\hat{C}(\boldsymbol{\phi})}{\hat{D}(\boldsymbol{\phi})} = \frac{C(\boldsymbol{\phi})}{D(\boldsymbol{\phi})} + \frac{1}{D(\boldsymbol{\phi})} \left\{ (\hat{C}(\boldsymbol{\phi}) - C(\boldsymbol{\phi})) - \frac{C(\boldsymbol{\phi})}{D(\boldsymbol{\phi})}(\hat{D}(\boldsymbol{\phi}) - D(\boldsymbol{\phi})) \right\} + o_p(n^{-1/2}).$$

Therefore, it holds that

$$\begin{aligned}\hat{\mathbf{g}}^*(\boldsymbol{\phi}; \mathbf{x}) &= \mathbf{g}^*(\boldsymbol{\phi}; \mathbf{x}) + \frac{(nh^d)^{-1}}{D(\boldsymbol{\phi}; \mathbf{x})} \sum_{j=1}^n r_j K_h(\mathbf{x} - \mathbf{x}_j) \pi^{-1}(\boldsymbol{\phi}; \mathbf{x}, y_j) O(\boldsymbol{\phi}; \mathbf{x}, y_j) \\ &\quad \times \{\mathbf{s}_0(\boldsymbol{\phi}; \mathbf{x}, y_j) - \mathbf{g}^*(\mathbf{x})\} + o_p(n^{-1/2}).\end{aligned}$$

Thus,

$$\begin{aligned}
& n^{-1} \sum_{i=1}^n \{1 - r_i/\pi(\boldsymbol{\phi}; \mathbf{z}_i)\} \hat{\mathbf{g}}^*(\boldsymbol{\phi}; \mathbf{x}_i) \\
&= n^{-1} \sum_{i=1}^n \{1 - r_i/\pi(\boldsymbol{\phi}; \mathbf{z}_i)\} \mathbf{g}^*(\boldsymbol{\phi}; \mathbf{x}_i) \\
&\quad + n^{-2} \sum_{i \neq j} \left[\frac{(1 - r_i/\pi(\boldsymbol{\phi}; \mathbf{z}_i)) r_j h^{-d} K_h(\mathbf{x}_i - \mathbf{x}_j) \pi^{-1}(\mathbf{x}_i, y_j) O(\boldsymbol{\phi}; \mathbf{x}_i, y_j)}{D(\boldsymbol{\phi}; \mathbf{x}_i)} \right. \\
&\quad \left. \times \{s_0(\boldsymbol{\phi}; \mathbf{x}_i, y_j) - \mathbf{g}^*(\boldsymbol{\phi}; \mathbf{x}_i)\} \right] + o_p(n^{-1/2}) \\
&= n^{-1} \sum_{i=1}^n \{1 - r_i/\pi(\boldsymbol{\phi}; \mathbf{z}_i)\} \mathbf{g}^*(\boldsymbol{\phi}; \mathbf{x}_i) + \binom{n}{2}^{-1} \sum_{i < j} [\zeta_{ij} + \zeta_{ji}]/2 + o_p(n^{-1/2}),
\end{aligned}$$

where

$$\zeta_{ij} := \frac{(1 - r_i/\pi(\boldsymbol{\phi}; \mathbf{z}_i)) r_j h^{-d} K_h(\mathbf{x}_i - \mathbf{x}_j) \pi^{-1}(\boldsymbol{\phi}; \mathbf{x}_i, y_j) O(\boldsymbol{\phi}; \mathbf{x}_i, y_j) \{s_0(\boldsymbol{\phi}; \mathbf{x}_i, y_j) - \mathbf{g}^*(\boldsymbol{\phi}; \mathbf{x}_i)\}}{D(\boldsymbol{\phi}; \mathbf{x}_i)}.$$

Let $\mathbf{w} = (\mathbf{x}^\top, y, r)^\top$ and $h(\mathbf{w}_i, \mathbf{w}_j) := (\zeta_{ij} + \zeta_{ji})/2$. According to U-statistic Theory (van der Vaart 1998), we have

$$\binom{n}{2}^{-1} \sum_{i < j} h(\mathbf{w}_i, \mathbf{w}_j) = \frac{2}{n} \sum_{i=1}^n E\{h(\mathbf{w}_i, \mathbf{w}_j) \mid \mathbf{w}_i\} + o_p(n^{-1/2}).$$

We first show that $E(\zeta_{ij} \mid \mathbf{w}_i) = 0$. It holds because

$$\begin{aligned}
& E[R\pi^{-1}(\boldsymbol{\phi}; \mathbf{x}_i, Y) O(\boldsymbol{\phi}; \mathbf{x}_i, Y) \{s_0(\boldsymbol{\phi}; \mathbf{x}_i, Y) - \mathbf{g}^*(\boldsymbol{\phi}; \mathbf{x}_i)\} \mid \mathbf{x}_i] \\
&= f_1(\mathbf{x}_i) [E_1\{(\pi^{-1}(\boldsymbol{\phi}; \mathbf{x}_i, Y) O(\boldsymbol{\phi}; \mathbf{x}_i, Y) - \pi^{-1}(\boldsymbol{\phi}; \mathbf{x}_i, Y) O(\boldsymbol{\phi}; \mathbf{x}_i, Y)) \mathbf{g}^*(\boldsymbol{\phi}; \mathbf{x}_i) \mid \mathbf{x}_i\}] \\
&= \mathbf{0}.
\end{aligned} \tag{A.9}$$

For $E(\zeta_{ji} \mid \mathbf{w}_i)$, it does not vanish, in fact,

$$\begin{aligned}
& E(\zeta_{ji} \mid \mathbf{w}_i) \\
&= r_i \int \left\{ 1 - E \left(\frac{\pi(\phi_0; \mathbf{Z})}{\pi(\phi; \mathbf{Z})} \mid \mathbf{x} \right) \right\} \frac{\pi^{-1}(\phi; \mathbf{x}, y_i) O(\phi; \mathbf{x}, y_i) \{s_0(\phi; \mathbf{x}, y_i) - \mathbf{g}^*(\phi; \mathbf{x})\}}{E_1\{\pi^{-1}(\mathbf{Z}; \phi) O(\mathbf{Z}; \phi) \mid \mathbf{x}\} f(\mathbf{x}) P(R=1 \mid \mathbf{x})} \\
&\quad \times h^{-d} K_h(\mathbf{x} - \mathbf{x}_i) f(\mathbf{x}) d\mathbf{x} \\
&= r_i \mathbf{G}(\mathbf{z}_i; \phi) + O(h^{2m}),
\end{aligned}$$

where $\mathbf{G}(\mathbf{z}_i; \phi) = G_1(\mathbf{x}_i; \phi) \mathbf{G}_2(\mathbf{z}_i; \phi)$ and

$$\begin{aligned}
G_1 &= 1 - E \left(\frac{\pi(\phi_0; \mathbf{Z})}{\pi(\phi; \mathbf{Z})} \mid \mathbf{x}_i \right), \\
\mathbf{G}_2 &= \frac{\pi^{-1}(\phi; \mathbf{z}_i) O(\phi; \mathbf{z}_i) \{s_0(\phi; \mathbf{z}_i) - \mathbf{g}^*(\phi; \mathbf{x}_i)\}}{E_1\{\pi^{-1}(\mathbf{Z}; \phi) O(\mathbf{Z}; \phi) \mid \mathbf{x}_i\} P(R=1 \mid \mathbf{x}_i)}
\end{aligned}$$

By the regularity condition (C14), $O(h^{2m}) = O(n^{-1/2})$. As a result, our estimating equation can be asymptotically expanded as

$$n^{-1} \sum_{i=1}^n \left\{ \left(1 - \frac{r_i}{\pi(\phi; \mathbf{z}_i)} \right) \mathbf{g}^*(\phi; \mathbf{x}_i) + r_i \mathbf{G}(\mathbf{z}_i; \phi) \right\} + o_p(n^{-1/2}). \quad (\text{A.10})$$

It seems that the asymptotic variance may increase due to the second term $r\mathbf{G}(\phi)$, but this solution also attains the lower bound. Once we got an unbiased estimating equation $\sum_{i=1}^n \varphi(\mathbf{z}_i; \phi) = 0$, the asymptotic variance would be given as $\text{Var}\{E(\dot{\varphi}(\phi_0))^{-1} \varphi(\phi_0)\}$, where $\dot{\varphi}(\phi_0) = \partial \varphi(\phi_0) / \partial \phi^\top$. Thus, to prove the fact, it is enough to show that $\mathbf{G}(\phi_0) = \mathbf{0}$ and $E(R\dot{\mathbf{G}}(\phi_0)) = \mathbf{0}$. The former equation is trivial, so we only show the latter equation. The latter equation can be written as $E(R\dot{\mathbf{G}}(\phi_0)) = E(RG_1(\phi_0)\dot{\mathbf{G}}_2(\phi_0)) + E(R\mathbf{G}_2(\phi_0)\dot{G}_1(\phi_0))$. It follows the first term is zero from $G_1(\phi_0) = 0$. Also, the second term is $E(R\mathbf{G}_2(\phi_0)\dot{G}_1(\phi_0)) = E\{E(R\mathbf{G}_2(\phi_0) \mid \mathbf{X})\dot{G}_1(\phi_0)\} = \mathbf{0}$. The last equation holds by the same reason as (A.9). Therefore, $r\mathbf{G}(\phi)$ does not effect on the asymptotic variance and our estimator also attains the semiparametric efficiency bound. \square

Appendix C

Riddles' estimator and our proposed estimator

Assume that X_1 , X_2 , and Y are binary estimators of the response probability π_{x_1y} (fully nonparametric model) and that they do not exceed unity. Let us denote the total number of each element of $\{i \mid R_i = 0, X_{1i} = a, X_{2i} = b, Y_i = c\}$, $\{i \mid R_i = 1, X_{1i} = a, X_{2i} = b, Y_i = c\}$, and $\{X_{1i} = a, X_{2i} = b, Y_i = c\}$ by ℓ_{abc} , m_{abc} , and n_{abc} , respectively. Note that ℓ_{abc} is unobservable, but $\ell_{ab\cdot} = \ell_{ab0} + \ell_{ab1}$ can be observed. We will consider two estimators: that of Riddles et al. (2016) and our proposed estimator. Riddles' estimator is a solution of

$$\sum_{i=1}^n \left\{ r_i \frac{(I_{00i}, I_{01i}, I_{10i}, I_{11i})^\top}{\pi_{x_{1i}y_i}} - (1 - r_i) \hat{E}_0 \left(\frac{(I_{00i}, I_{01i}, I_{10i}, I_{11i})^\top}{1 - \pi_{x_{1i}Y}} \mid x_{1i}, x_{2i} \right) \right\} = 0,$$

where $I_{aci} = I(X_{1i} = a, Y_i = c)$ for $a, c = 0, 1$, and $\hat{E}_0(\cdot \mid x_1, x_2)$ is a nonparametric estimator of the expectation $E_0(\cdot \mid x_1, x_2)$. By using the same idea as in (15), it follows that if $x_{1i} = \alpha$ and $x_{2i} = \beta$,

$$\hat{E}_0 \left(\frac{I_{aci}}{1 - \pi_{\alpha Y}} \mid x_1 = \alpha, x_2 = \beta \right) = I(a = \alpha) \frac{m_{\alpha\beta c} / \pi_{\alpha c}}{\sum_{k=0}^1 m_{\alpha\beta k} (1 - \pi_{\alpha k}) / \pi_{\alpha k}}.$$

Thus, the equation for I_{ac} can be simplified to

$$\frac{m_{a\cdot c}}{\pi_{ac}} - \sum_{\beta=0}^1 \ell_{a\beta\cdot} \frac{m_{a\beta c} / \pi_{ac}}{\sum_{k=0}^1 m_{a\beta k} (1 - \pi_{ak}) / \pi_{ak}} = 0, \quad (\text{A.11})$$

where $m_{a\cdot c} = m_{a0c} + m_{a1c}$.

Next, consider our proposed estimator, which is a solution of

$$\sum_{i=1}^n \left(1 - \frac{r_i}{\pi_{x_{1i}y_i}} \right) \hat{E}^\star \left(\frac{(I_{00i}, I_{01i}, I_{10i}, I_{11i})^\top}{1 - \pi_{x_{1i}Y}} \mid x_{1i}, x_{2i} \right) = 0.$$

By using the same arguments as we used above, this can be simplified to

$$\sum_{\beta=0}^1 \left(n_{a\beta} - \frac{m_{a\beta 0}}{\pi_{a0}} - \frac{m_{a\beta 1}}{\pi_{a1}} \right) \frac{m_{a\beta c}/\pi_{ac}^2}{\sum_{k=0}^1 m_{a\beta k}(1 - \pi_{ak})/\pi_{ak}^2} = 0. \quad (\text{A.12})$$

We will now show that (A.11) and (A.12) have the same solution. We first solve (A.12). For $a = 0, 1$, from (A.12), we have that π_{a0} and π_{a1} are solutions of

$$\begin{aligned} n_{a0} - m_{a00}/\pi_{a0} - m_{a01}/\pi_{a1} &= 0, \\ n_{a1} - m_{a010}/\pi_{a0} - m_{a11}/\pi_{a1} &= 0, \end{aligned} \quad (\text{A.13})$$

which can be easily solved. Next, by (A.11), we have that π_{a0} and π_{a1} are solutions of the following simultaneous equations: $m_{a00}P_0 + m_{a10}P_1 = m_{a.0}$ and $m_{a01}P_0 + m_{a11}P_1 = m_{a.1}$, where $P_0 = \ell_{a0}(\sum_{k=0}^1 m_{a0k}(1 - \pi_{ak})/\pi_{ak})^{-1}$ and $P_1 = \ell_{a1}(\sum_{k=0}^1 m_{a1k}(1 - \pi_{ak})/\pi_{ak})^{-1}$. We can solve this to obtain $P_0 = P_1 = 1$. In fact, for ξ_0 , $(m_{a01} - m_{a00}m_{a01}^{-1}m_{a11})P_0 = m_{a.1} - m_{a.0}m_{a01}^{-1}m_{a11} = m_{a01} - m_{a00}m_{a01}^{-1}m_{a11}$, and the same holds for P_1 . From this, we have $\sum_{k=0}^1 m_{a0k}(1 - \pi_{ak})/\pi_{ak} = \ell_{a0}$. and $\sum_{k=0}^1 m_{a1k}(1 - \pi_{ak})/\pi_{ak} = \ell_{a1}$. These two equations imply (A.13), because $\ell_{aj} = n_{aj} - m_{aj0} - m_{aj1}$ for $j = 0, 1$. Therefore, (A.11) and (A.12) have the same solution.

Qin et al. (2002)'s and Chang & Kott (2008)'s estimator

Let $\pi(\mathbf{z}; \boldsymbol{\phi})$ be a response mechanism, known up to $\boldsymbol{\phi} \in \Phi \subset \mathbb{R}^p$, and let it be a logistic model: $\pi(\mathbf{z}; \boldsymbol{\phi}) = [1 + \exp\{h(\mathbf{z}; \boldsymbol{\phi})\}]^{-1}$. Suppose that for the following equation, there exists a function $\mathbf{f} : \mathcal{X} \rightarrow \mathbb{R}^{p-1}$ and a unique solution $\hat{\boldsymbol{\phi}}$:

$$\sum_{i=1}^n \left(1 - \frac{r_i}{\pi(\mathbf{z}_i; \boldsymbol{\phi})} \right) \mathbf{g}(\mathbf{x}_i) = \mathbf{0}, \quad (\text{A.14})$$

where $\mathbf{g}(\mathbf{x}) = (1, \mathbf{f}(\mathbf{x})^\top)^\top$. Then, the solution is exactly the same as that in Qin et al. (2002) that is constrained by $\omega_i \geq 0$, $\sum_{i=1}^n r_i \omega_i = 1$, $\sum_{i=1}^n r_i \omega_i (\pi(\mathbf{z}_i; \boldsymbol{\phi}) - W)$, and

$$\sum_{i=1}^n r_i \omega_i \mathbf{g}(\mathbf{x}_i) = n^{-1} \sum_{i=1}^n \mathbf{g}(\mathbf{x}_i) =: \bar{\mathbf{g}}.$$

Let $d_i(\lambda_1, W, \phi) := 1 + \boldsymbol{\lambda}_1^\top (\mathbf{g}(\mathbf{x}_i) - \bar{\mathbf{g}}) + \lambda_2 (\pi(\mathbf{z}_i; \boldsymbol{\phi}) - W)$. According to (7)–(10) in Qin et al. (2002), the estimator is a solution of the following simultaneous equations:

$$\sum_{i=1}^n r_i \frac{\mathbf{g}(\mathbf{x}_i) - \bar{\mathbf{g}}}{d_i(\boldsymbol{\lambda}_1, W, \phi)} = \mathbf{0}, \quad (\text{A.15})$$

$$\sum_{i=1}^n r_i \frac{\pi(\mathbf{z}_i) - W}{d_i(\boldsymbol{\lambda}_1, W, \phi)} = 0, \quad (\text{A.16})$$

$$\sum_{i=1}^n r_i \frac{\partial \log \pi(\mathbf{z}_i; \boldsymbol{\phi})}{\partial \boldsymbol{\phi}} - \lambda_2 \sum_{i=1}^n r_i \frac{\partial \pi(\mathbf{z}_i; \boldsymbol{\phi}) / \partial \boldsymbol{\phi}}{d_i(\boldsymbol{\lambda}_1, W, \phi)} = \mathbf{0}, \quad (\text{A.17})$$

$$\lambda_2 = (n/n_r - 1)/(1 - W), \quad (\text{A.18})$$

where $n_r = \sum_{i=1}^n r_i$. Note that the parameters to be estimated are $\boldsymbol{\lambda}_1$, W , and $\boldsymbol{\phi}$.

It follows from (A.17) and the assumption that the response model is logistic that

$$\begin{aligned} & \sum_{i=1}^n r_i \{1 - \pi(\mathbf{z}_i; \boldsymbol{\phi})\} \dot{\mathbf{h}}(\mathbf{z}_i; \boldsymbol{\phi}) \\ &= \lambda_2 \sum_{i=1}^n r_i \frac{\pi(\mathbf{z}_i; \boldsymbol{\phi}) \{1 - \pi(\mathbf{z}_i; \boldsymbol{\phi})\}}{d_i(\boldsymbol{\lambda}_1, W, \phi)} \dot{\mathbf{h}}(\mathbf{z}_i; \boldsymbol{\phi}). \end{aligned} \quad (\text{A.19})$$

Thus, if we set $\boldsymbol{\lambda}_1 = \mathbf{0}$, $W = n_r/n$, $\lambda_2 = n/n_r$, and $d_i(\boldsymbol{\lambda}_1, W, \phi) = \lambda_2 \pi(\mathbf{z}_i)$, this implies that (A.19) or (A.17) holds. By substituting these results into (A.15) and (A.16), we have (A.14). Note that we assumed that the equation has a unique solution and the probability that the estimator of Qin et al. (2002) has a unique solution goes to one as the sample size goes to infinity, and thus, this is the solution.

References

- Bickel, P. J., Klaassen, C. A. J., Ritov, Y. & Wellner, J. A. (1998), *Efficient and Adaptive Estimation for Semiparametric Models*, Springer.
- Chang, T. & Kott, P. S. (2008), ‘Using calibration weighting to adjust for nonresponse under a plausible model’, *Biometrika* **95**(3), 555–571.
- Diggle, P. & Kenward, M. G. (1994), ‘Informative drop-out in longitudinal data analysis’, *Journal of the Royal Statistical Society Ser. C* **43**(1), 49–93.
- Fitzmaurice, G. M., Lipsitz, S. R., Molenberghs, G. & Ibrahim, J. G. (2005), ‘A protective estimator for longitudinal binary data subject to non-ignorable non-monotone missingness’, *Journal of the Royal Statistical Society Ser. A* **168**(4), 723–735.
- Han, P. (2014), ‘Multiply robust estimation in regression analysis with missing data’, *Journal of the American Statistical Association* **109**(507), 1159–1173.
- Kim, J. K. (2011), ‘Parametric fractional imputation for missing data analysis’, *Biometrika* **98**(1), 119–132.
- Kim, J. K. & Yu, C. L. (2011), ‘A semiparametric estimation of mean functionals with non-ignorable missing data’, *Journal of the American Statistical Association* **106**(493), 157–165.
- Liang, K.-Y. & Zeger, S. L. (1986), ‘Longitudinal data analysis using generalized linear models’, *Biometrika* **73**(1), 13–22.
- Little, R. J. A. & Rubin, D. B. (2002), *Statistical Inference with Missing Data*, 2nd edition edn, Wiley Series in Probability and Statistics.

- Louis, T. A. (1982), ‘Finding the observed information matrix when using the em algorithm’, *Journal of the Royal Statistical Society Ser. B* **44**(2), 226–233.
- Miao, W. & Tchege Tchetgen, E. J. (2016), ‘On varieties of doubly robust estimators under missingness not at random with a shadow variable’, *Biometrika* **103**(2), 475–482.
- Morikawa, K., Kim, J. K. & Kano, Y. (2016), ‘Semiparametric maximum likelihood estimation under nonignorable nonresponse’, *Submitted*.
- Qin, J., Leung, D. & Shao, J. (2002), ‘Estimation with survey data under nonignorable nonresponse or informative sampling’, *Journal of the American Statistical Association* **97**(457), 193–200.
- Riddles, M. K., Kim, J. K. & Im, J. (2016), ‘Propensity-score-adjustment method for nonignorable nonresponse’, *Journal of Survey Statistics and Methodology* **4**(2), 215–245.
- Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994), ‘Estimation of regression coefficients when some regressors are not always observed’, *Journal of the American Statistical Association* **89**(427), 846–866.
- Rotnitzky, A. & Robins, J. (1997), ‘Analysis of semi-parametric regression models with non-ignorable non-response’, *Statistics in Medicine* **16**, 81–102.
- Rotnitzky, A., Scharfstein, D., Su, T.-L. & Robins, J. (2001), ‘Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring’, *Biometrics* **57**(1), 103–113.
- Scharfstein, D. O., Rotnizky, A. & Robins, J. M. (1999), ‘Adjusting for nonignorable drop-out using semiparametric nonresponse models’, *Journal of the American Statistical Association* **94**, 1096–1146.

- Shao, J. & Wang, L. (2016), ‘Semiparametric inverse propensity weighting for nonignorable missing data’, *Biometrika* **103**(1), 175–187.
- Skrondal, A. & Rabe-Hesketh, S. (2014), ‘Protective estimation of mixed-effects logistic regression when data are not missing at random’, *Biometrika* **101**(1), 175–188.
- Sverchkov, M. (2008), ‘A new approach to estimation of response probabilities when missing data are not missing at random’, *In Proc. Survey Res. Meth. Sect., Am. Statist. Assoc.* Washington DC: American Statistical Association, 867-874 .
- Tang, G., Little, R. J. A. & Raghunathan, T. E. (2003), ‘Analysis of multivariate missing data with nonignorable nonresponse’, *Biometrika* **90**(4), 747–764.
- Tang, N., Zhao, P., Qu, A. & Jiang, D. (2016), ‘Semiparametric estimating equations inference with nonignorable nonresponse’, *Statistica Sinica* .
- Tang, N., Zhao, P. & Zhu, H. (2014), ‘Empirical likelihood for estimating equations with nonignorably missing data empirical likelihood for estimating equations with nonignorably missing data’, *Statistica Sinica* **24**(2), 723–747.
- Tsiatis, A. A. (2006), *Semiparametric Theory and Missing Data*, Springer Series in Statistics, Springer.
- van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge University Press.
- Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E. & Kenward, M. G. (2001), ‘Sensitivity analysis for nonrandom dropout: A local influence approach’, *Biometrics* **57**(1), 7–14.
- Wang, S., Shao, J. & Kim, J. K. (2014), ‘An instrumental variable approach for identification and estimation with nonignorable nonresponse’, *Statistica Sinica* **24**(3), 1097–1116.

Zhao, J. & Shao, J. (2015), ‘Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data’, *Journal of the American Statistical Association* **110**(512), 1577–1590.